	Analyzing SARS-CoV-2 Data in Terra using Theiagen's TheiaCoV FASTA Workflow Version 1		
	Document TG-SC2-FST, Version 2		
	Date:	Effective Date:	Workflow Version
	9/18/2023	9/2023	PHB v1

1. PURPOSE/SCOPE

To standardize the process of analyzing SARS-COV-2 (SC2) next generation sequencing (NGS) data using Theiagen's TheiaCoV_FASTA_PHB workflow in Terra to determine quality control (QC) metrics, Nextclade clade, and Pangolin lineage assignments. Acceptable data types include the FASTA file format.

2. REQUIRED RESOURCES

- Computer
- Internet connection: at least 10 and 5Mbps for download and upload speeds, respectively
- Internet browser
  - Google Chrome, Firefox, or Edge
- Google account
- Terra account, linked to Google account
- FASTA files uploaded to Terra workspace, see TG-TER-03
- Theiagen's TheiaCoV_FASTA_PHB workflow in Terra, see TG-TER-03 appendix 9.2

IMPORTANT NOTES

- Metadata column headers and workflow input text indicated in gray in this SOP are customizable; black is required text
- Terra data table column headers become available as workflow inputs when running workflows, search for them in workflow input dropdowns using the prefix *this.* to filter
- Filter for workspace data and files in workflow input dropdowns using the prefix *workspace.*

3. RELATED DOCUMENTS

Document Number	Document Name
TG-TER-03	Uploading Local or SRA NGS Data & Creating a Results Metadata Table in Terra

4. PROCEDURE

4.1 CREATE A SAMPLE METADATA FILE (TSV FILE) FOR ASSEMBLIES

1. In Excel, *create a list* containing the following sample information (Fig 1):
   a. Column 1 header: *entity:FASTA_Test_id*, where *FASTA_Test* is the data table/group of samples to be analyzed
   b. List all *sample IDs* in column 1
   c. Column 2 header: *assembly_fasta*, or similar
   d. *Optional: remaining columns may be used to add metadata like additional lab results, sample collection information, demographic data, etc*

entity:FASTA_Test_id	assembly_fasta	run_id
Sample_01	gs://sc2_validato	SEQ197
Sample_02	gs://sc2_validato	SEQ197
	validato	SEQ197

Figure 1: Assembly Metadata file.

   e. Do not include spaces in the headers
2. *Save as* a txt or tsv file
3. *Upload* to Terra workspace; see TG-TER-03 for details

4.2 RUNNING THE THEIACOV WORKFLOW

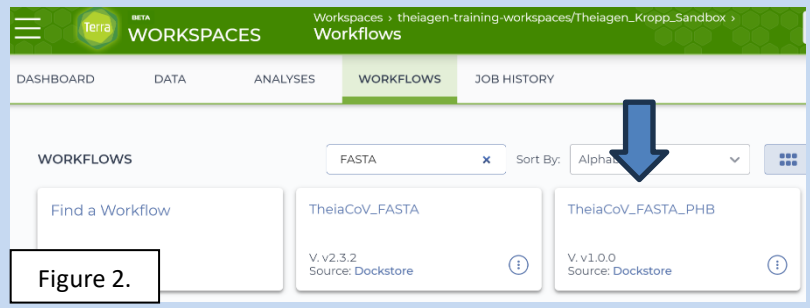

Figure 2.

1. Open Terra and navigate to the *workflows* tab within the workspace containing SC2 data
2. Select the *TheiaCoV_FASTA_PHB* (Fig 2)
3. *Uncheck call caching* (Fig 3)
4. Choose the latest version of *version 1*, or the version used for internal validation (Fig 3, a)
5. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 3, b)
6. Select the relevant data table under the select *root entity type* dropdown (Fig 3, c)
7. Click *select data* (Fig 3, d)

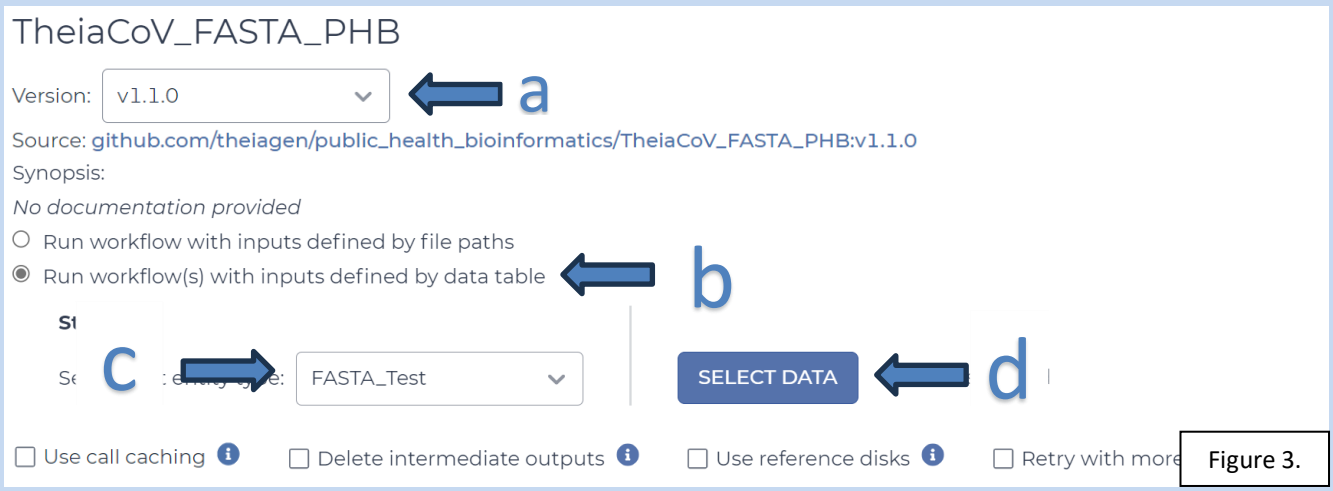

Figure 3.

8. In the pop-up window *select the checkbox* for each sample to include in the analysis (Fig 4)
   - *a* *Click the down arrow and select all to process all specimens*
   - *b* *Additionally, a subset of samples may be chosen using the search bar to filter before selecting the checkbox at the top to only select samples matching the search criteria*
   - c Scroll to the bottom and click *ok*

9. On the inputs tab, *upload the TheiaCov input json file*; for the newest version, navigate to the Theiagen Public Health Resources page at https://theiagen.notion.site/Theiagen-Public-Health-Resources-a4bd134b0c5c4fe39870e21029a30566 and click the first link in the Key Resources box titled Docker Image and Reference Materials for SARS-CoV-2 Genomic Characterization
   a Scroll down and expand the *Terra.Bio Input JSONs*; click on the json file associated with FASTA files, TheiaCoV_FASTA_PHB_2023-08-24.json file (the date may vary to reflect the most up-to-date version)
   b *Right click* and *save* the file (text does not have to be selected to save properly)
10. Return to the workflow in Terra, click *upload json* (Fig 5, red circle), *select* the saved json file, and click *open*
   a This will set the dataset tags and docker image components of the workflow

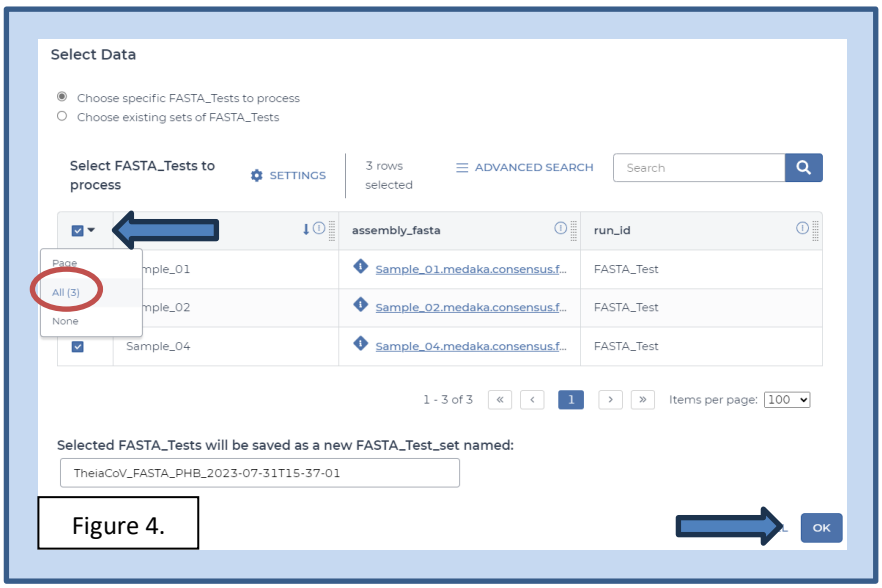

Figure 4.

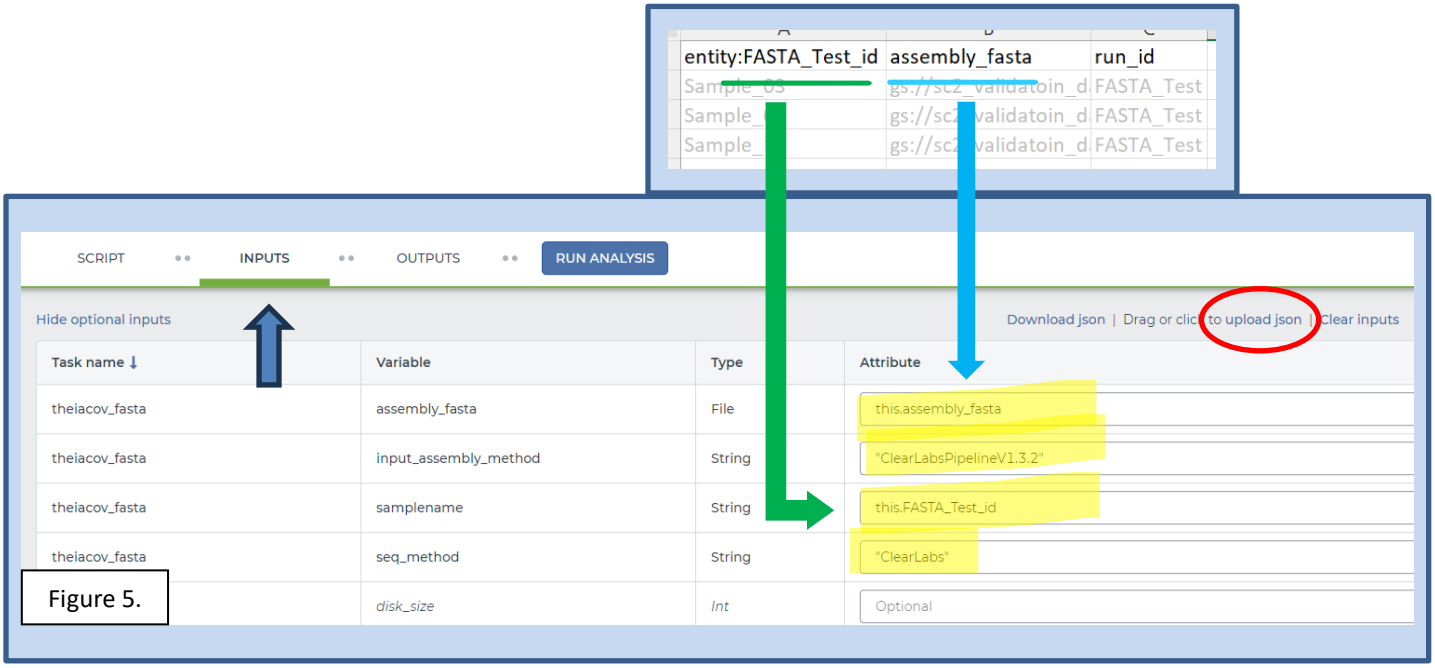

Figure 5.

11. Set the first four attributes in the table to the following, respectively (Fig 5):
    a *this.assembly_fasta* (must match unique column 2 header text in tsv file)
    b *"ClearLabsPipelineV1.3.2"* (the relevant assembly pipeline in "quotes")
    c *this.FASTA_Test_id* (must match unique column 1 header text in tsv file)
    d *"ClearLabs"* (the relevant sequencing method in "quotes")
12. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 6)
13. Click *save*
14. Launch the workflow by clicking *run analysis* (Fig 6); enter desired comments and click *launch*

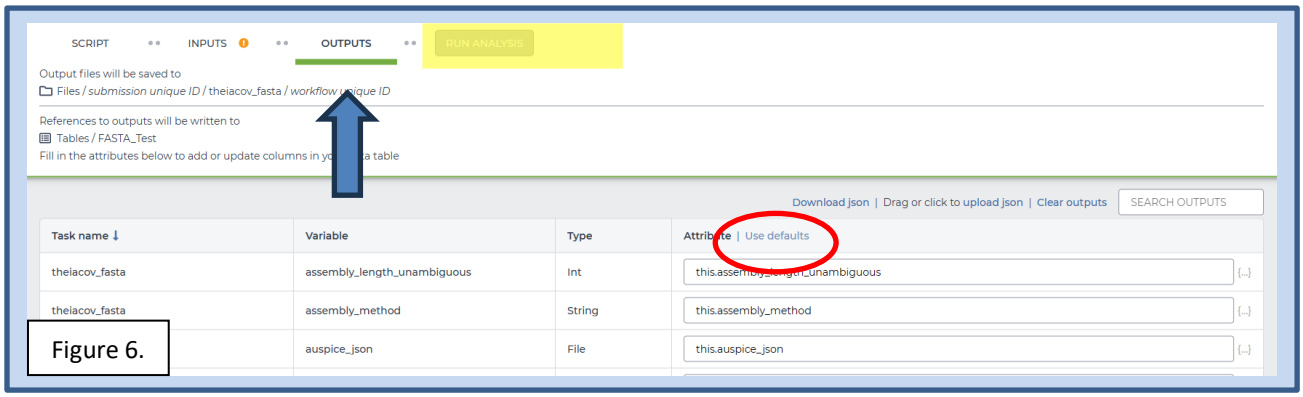

Figure 6.

4.3 QUALITY ASSESSMENT OF THEIACOV OUTPUTS

1. Navigate to the *data* tab of the workspace containing SC2 data and open the pertinent data table
2. Click *settings* (Fig 7, green rectangle) and select *none* to deselect all output columns (Fig 8, yellow highlight)

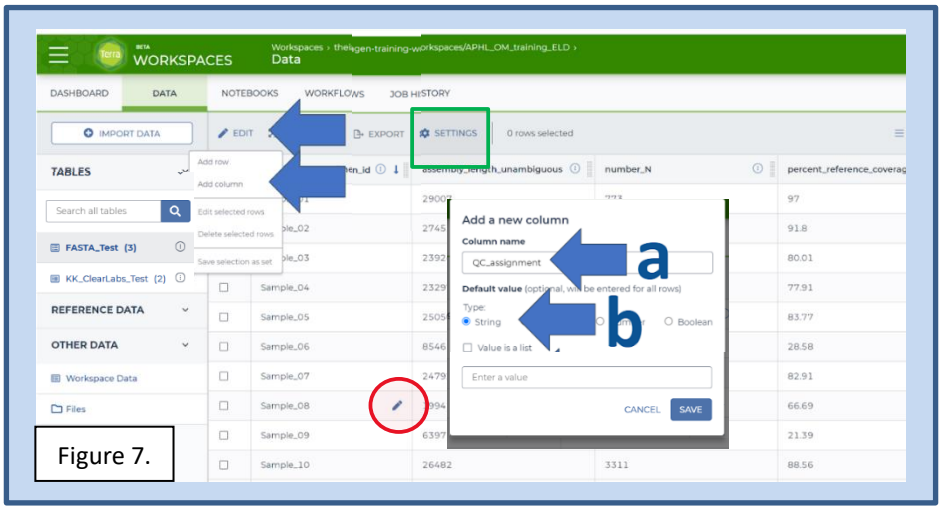

Figure 7.

3. To simplify the table, select the three following outputs that will be used to make a QC assessment: *assembly_length_unambiguous*, *Number_N*, and *percent_reference_coverage*
   a. *Optional: save this selection by clicking in the* save this column selection *field and* naming it *(e.g. QC_assessment); do not include any spaces in the name (Fig 8, red rectangle)*
   b. Click *done*

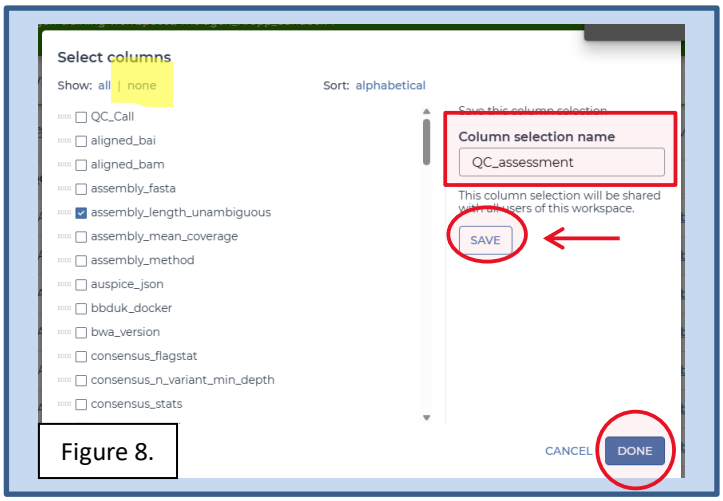

Figure 8.

4. *Optional: add a column to record QC PASS/FAIL by clicking* edit, add a column *(Fig 7)*
   a. *Name the new column (e.g. QC_Call); do not include any spaces*
   b. *Set the value type as* string
   c. *Click* save
5. Use table 1 to assess the quality of each sample's genome assembly (see next page) &/or lab-specific quality metrics

Table 1. Guidance thresholds for genome assembly QC

QC Metric	Guidance Threshold*[1]
Number N	<5kbp
Assembly length unambiguous	>24kbp
Percent reference coverage	>83%

6. *Optional: notate in the QC_assessment field for each sample* PASS *or* FAIL *by* clicking the pencil icon *in the corresponding field (Fig 7, red circle)*
7. For samples that pass the guidance thresholds, proceed to section 4.4
   a. For samples that do not pass guidance thresholds, resequence
      i. Samples not meeting guidance thresholds indicated here may proceed to analysis at the discretion of the laboratory

[1] Metrics and thresholds are presented for guidance only as there are currently no standard assembly metric requirements; internal validation procedures will ultimately define acceptable assembly QC parameters

4.4 DETERMINING SARS-CoV-2 CLADES, LINEAGES, AND WHO VARIANTS OF CONCERN (VoC)

1. Navigate to the *data* tab of the Terra workspace containing SC2 data of interest
2. *Open the data table* by clicking on the name of the data table in the left sidebar
3. View *settings* above the data table (Fig 7), select *none* (Fig 8)
4. Select the following columns: *nextclade_clade* and *pango_lineage*
   a. Save this column group for future use by clicking the *save this column selection* field, *naming it* (e.g. SC2_Results), and clicking *save*
5. Click *done*
6. Determine the Nextclade clade for each sample
   a. In the data table, find the column titled *nextclade_clade*; result formats will use the following nomenclature: *21L (Omicron)* where:
      i. *21L* indicates the sample clade and
      ii. In parentheses, *(Omicron)*, contains the WHO variant of concern classification
         1. *Not every sample will belong to a WHO classification*
   b. *Samples indicated as recombinant may indicate a case where multiple strains have combined during viral replication producing a new lineage*
   c. *More information on SARS-CoV-2 recombinants can be found at the following Github site: pipeline-resources/docs/sc2-recombinants.md at main · pha4ge/pipeline-resources · GitHub*
7. Identify the Pangolin lineage for each sample
   a. In the data table, find the column titled *pango_lineage*; nomenclature will be similar to the following: B.1.167
   b. *For more information on each of the lineages, visit https://cov-lineages.org/lineage_list.html*
8. Follow lab-specific QC, resulting, and reporting procedures, as applicable

5. QUALITY RECORDS

- Raw read files
- Sample read and assembly QC metrics
- All workflow outputs relevant to results, including tool and database versions

6. TROUBLESHOOTING

- Consult with internal staff familiar with this procedure or contact support@theiagen.com for troubleshooting inquiries
- For document edit requests, contact support@theiagen.com

7. INTERFERENCES

N/A

8. REFERENCES

1. Smith, E., Wright, S., & Libuit, K. (2022, June 28). *Identifying SARS-CoV-2 Recombinants*. Github. Retrieved June 16, 2023, from https://github.com/pha4ge/pipeline-resources/blob/main/docs/sc2-recombinants.md#identifying-sars-cov-2-recombinants

2. O'Toole, Áine et al. "Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch." *Wellcome open research* vol. 6 121. 17 Sep. 2021, doi:10.12688/wellcomeopenres.16661.2

9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	7/2023
Uncheck call caching, updated input json, figures, and formatting	2	9/2023

10. APPENDICES

None